Optimal analysis of complex protein mass spectra

Dijkstra, Martijn; Jansen, Ritsert C.

Published in:
Proteomics

DOI:
10.1002/pmic.200701064

Publication date:
2009

RESEARCH ARTICLE

Optimal analysis of complex protein mass spectra

Martijn Dijkstra[1,2] *and Ritsert C. Jansen*[1,3]

[1] Groningen Bioinformatics Centre, University of Groningen, Haren, The Netherlands
[2] Center for Medical Biomics, University of Groningen, Groningen, The Netherlands
[3] Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands

Received: November 19, 2007
Revised: May 5, 2009
Accepted: May 6, 2009

Due to physical and chemical phenomena, a simple sample can give rise to a complex mass spectrum with many more peaks than the number of molecular species present in the sample. We link peaks within and between different spectra, and come up with an advanced analysis approach to produce reliable estimates of the molecule masses and abundances. By linking peaks, we can locate multiple-charge peaks at the correct position in the spectrum, we can deconvolute complex regions with many overlapping peaks by including information from related regions with lower complexity and higher resolution, and we reduce the total number of observed peaks in a spectrum to a much smaller number of underlying molecular species. In this paper we properly model 29 952 peaks in 64 spectra, using only 39 location parameters and one shape parameter. This major reduction from many different molecules to a limited set of molecular species reduces the statistical test multiplicity for biomarker discovery and therefore we imply that the reduction should eventually increase the biomarker discovery power significantly, too.

Keywords:
Biomarker discovery / Calibration / Deconvolution / MS / Mixture models

1 Introduction

A simple sample containing few molecular species can generate a complex mass spectrum with many peaks. Various chemical and physical phenomena can explain this [1]. For example, a molecular species can have different forms (isotopes) with different numbers of neutrons. These isotopes give rise to peaks at multiple locations $\mu+n$ in the spectrum (mono-isotopic molecule mass μ with $n = 0,1,2,\ldots$ neutrons). High-resolution mass spectrometers can detect isotopes as separate peaks in the spectrum. In addition, molecules can get different numbers of charges. The number of charges that a molecule generally gets depends on the character of the molecule and also on the ionization technique used. The number ranges from one to three for SELDI and MALDI, to over 30 for ESI. For each charge state ($z = 1,2,3,\ldots$), there will be a series of peaks in the spectrum. Molecules of a given molecular species can also form intermolecular complexes, for example with zero or more ($a = 0,1,2,3\ldots$) matrix molecules in SELDI and MALDI. This also gives rise to multiple peaks at locations $(\mu+n+a\cdot\mu_a)/z$ in the spectrum (adduct mass μ_a). In this way, the combination of variable numbers of neutrons, charges and matrix adducts can give rise to a multitude of peaks *per* molecular species. Strikingly, current statistical methods for calibration and analysis of mass spectra (*e.g.* [2–7]) do not exploit this interconnectivity between peaks and instead treat all peaks as independent species. At best, [8] suggest to superimpose plots of the spectrum against m/z and of the spectrum against $2\times m/z$, as a quick check of whether the data was calibrated appropriately; the single- and double-charge peaks should line up. In this paper, we present new and improved methods to link peaks within a spectrum and across different spectra. We anticipate that the new approach offers a number of advantages. First, a spectrum can be ''self-calibrated'' so that multiple-charge peaks locate at the correct positions. Second, complex regions with many overlapping peaks can be deconvoluted by using information from related regions with lower complexity (*e.g.* double-

Correspondence: Dr. Martijn Dijkstra, Groningen Bioinformatics Centre, University of Groningen, Kerklaan 30, 9751NN Haren, The Netherlands
E-mail: m.dijkstra@rug.nl
Fax: +31-50-363-8971

Abbreviation: SPA, sinapinic acid

charge peaks have higher resolution than single-charge peaks and can therefore help to define the number of single-charge peaks). Third, the total number (say 1000) of observed peaks in a spectrum can be reduced to a much smaller number of underlying molecular species (say 100), which reduces the statistical test multiplicity in the biomarker discovery phase and therefore increases the power significantly. We demonstrate these properties, using SELDI-TOF MS data. Recently, SELDI-TOF MS was used in several biomarker discovery studies because it can generate hundreds of spectra *per* day by using high-throughput robot-automated sample preparation [9–11]. For a recent overview of SELDI-TOF's analytical opportunities and technical limitations we refer to [12]. We also discuss the application and benefits of our methods to other MS technologies. In SELDI or other MS analyzes, we and others do not have a ''gold standard'' data set providing us with complete *a priori* known specifications of all peak parameters (including peak heights) [13]. This implies that we can never unambiguously prove our approach.

2 Materials and methods

2.1 SELDI-TOF MS data

Figure 1 presents real data from serum samples, which were measured with a low-resolution SELDI-TOF mass spectrometer from Ciphergen. The serum samples were taken from patients treated for colon cancer; [14] give a detailed description of the samples. The ''SELDI method'' involves three steps: a specific fraction of molecules is enriched from the sample; the selected molecules are then embedded in a lattice of energy absorbing molecules (also known as matrix molecules); and the energy absorbing molecules use the energy from a laser to sublimate and ionize the selected molecules. The ''TOF method'' makes use of an electric field to separate and detect the charged molecules based on their m/z.

The upper panel in Fig. 1 apparently shows two peaks that correspond to two single-charged molecular species. These two peaks are skewed and show shoulders. This is explained by the formation of intermolecular complexes of sample molecules with 0, 1, 2 and 3 matrix adducts, which here leads to $2 \times 3 = 6$ extra peaks in the spectrum [1]. However, the extra peaks can hardly be seen; a simple deconvolution method would probably just fit two skewed distributions to the spectrum.

The complexes can also get 1, 2 or 3 charges. The lower panel in Fig. 1 shows that molecules with >1 charges generate peaks with higher resolution so that more peaks can be detected. Double- and triple-charge peaks can therefore provide helpful insight into the complexity of a mixture of single-charge peaks suffering from overlapping of peaks.

2.2 Calibration of TOF MS data

One is generally interested in the mass of the molecules and not in their TOF. Therefore, the processing of samples typically starts with a calibration run. The measured TOFs of molecules with known masses in a synthesized sample can be used to set the calibration parameters. This is generally done by estimating the parameters such that the sum of the squared differences between measured and predicted m/z's is minimal. These parameters are used for the conversion

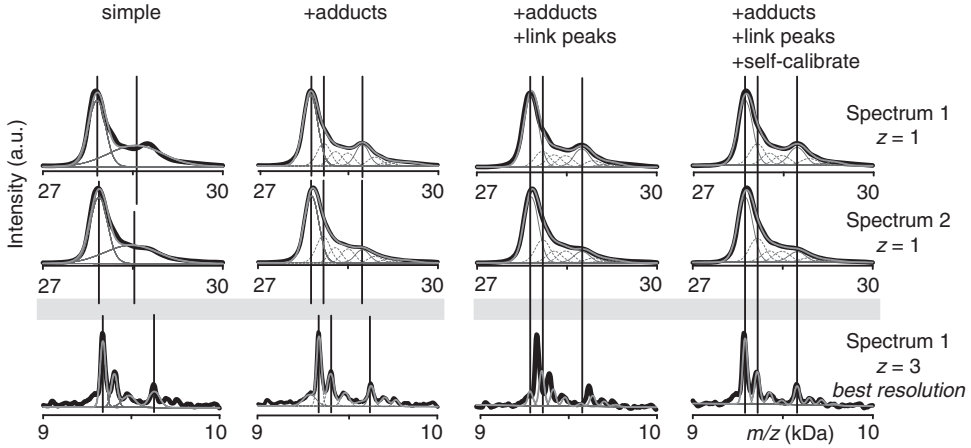

Figure 1. Rows 1 and 2 show the single-charge peaks (bold black curves) in Spectra 1 and 2, respectively. Row 3 shows the corresponding triple-charge peaks in Spectrum 1. Every column shows a fit of a mixture model (bold gray curves) to the data, ranging from a simple model in the left column to a complex/improved model in the right column, as indicated in the headings. The thin solid and dashed curves correspond to the individual mixture distributions; solid means 0 adducts, dashed means >0 adducts. The thin vertical black lines indicate the estimated peak locations. Incorporating adducts in the model improves the fit, and linking peaks reduces the total number of parameters. Self-calibration reduces the misalignment between corresponding peaks within the spectrum.

from TOF data to m/z data in the next runs. The derived m/z data may be displayed and analyzed visually and computationally.

Unfortunately, calibration parameters derived from one spectrum do not always apply well to other spectra, *i.e.* first, locations of corresponding peaks can be shifted across different spectra, and, second, within a single spectrum, double-charge peaks are not located exactly at half the mass position of the single-charge peaks. A reason can be a variable length of the flight tube. Peak shifts between spectra are small if the spectra are measured with a single instrument and within a short period of time [15]. Here, we propose to "*self-calibrate*" a spectrum and address the second issue so that multiple-charge peaks locate at the correct relative positions as compared with the single-charge peak. Our method can also be used to align different spectra.

A spectrum is most often visualized as a "smoothed" histogram of the detected intensities of molecules, with the horizontal axis on m/z-scale. We transform the TOF (t) on the horizontal axis to m/z-values (γ), by means of a quadratic calibration equation,

$$\frac{\gamma(t|\alpha, t_0, \beta)}{U} = \alpha(t - t_0)^2 + \beta \quad (1)$$

with calibration parameters, α, t_0 and β, and the known, applied electric field voltage U [16].

2.3 Self-calibration of TOF MS data

A given spectrum can be self-calibrated by determining optimal values for calibration parameters t_0 and β. The values are optimal if the locations of the double-charge peaks (z_2 peaks) in the spectrum best match the locations of the corresponding single-charge peaks (z_1 peaks) in the spectrum, as illustrated in Fig. 2. We use the correlation between the measured intensities on the normal m/z-axis and the measured intensities in the same spectrum at $2 \times m/z$, to indicate the goodness of a match. Technically, we search calibration parameters such that the locations of the z_1 peaks in region 1, *i.e.*

$$\left[2 \times \gamma_{\text{left}}, \gamma_{\text{right}}\right] \quad (2)$$

best match the locations of the z_2 peaks in the region two, *i.e.*

$$\left[\gamma_{\text{left}}, \frac{\gamma_{\text{right}}}{2}\right] \quad (3)$$

where γ_{left} and γ_{right} are two locations on the m/z-axis in the spectrum, *e.g.* the boundaries of the spectrum.

First, we linearly interpolate the intensities in region 1, at twice the m/z-values in region 2. Next we calculate the "*charge-correlation*", *i.e.* the correlation between the interpolated intensities in region 1 and the intensities in region 2. We use well-known methods to optimize the charge-correlation as function of the parameters t_0 and β. Prior baseline subtraction is recommended for spectra that suffer severely from chemical noise.

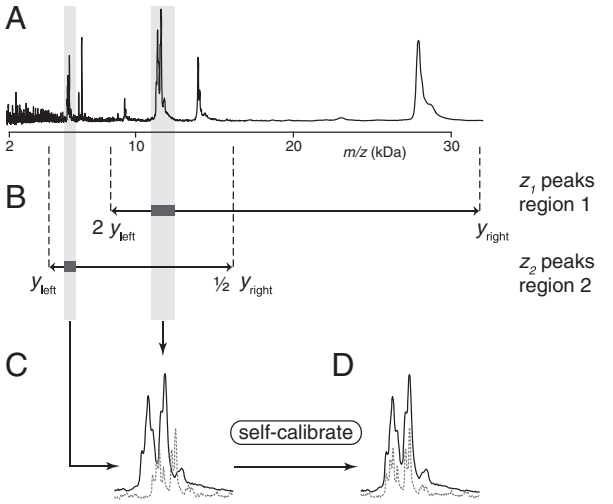

Figure 2. A detailed illustration of self-calibration of the spectrum shown in (A). Given two locations, γ_{left} and γ_{right}, on the m/z-axis, regions 1 and 2 are defined as shown in (B). Region 1 contains z_1 peaks (single-charge), which correspond to z_2 peaks (double-charge) in region 2. However, the relative locations of the z_1 and the z_2 peaks do not generally exactly match. (C) illustrates this in closeup by doubling the m/z locations of z_2 peaks and plotting them on top of the corresponding z_1 peaks. Self-calibration optimizes the correlation between the intensities in the two regions as function of the calibration parameters. As a result, the locations of the z_2 peaks match the locations z_1 peaks, as is shown in (D).

By finding optimal values for the parameters, t_0 and β, in the calibration Eq. 1, we locate peaks at their correct relative location in the spectrum. Therefore, if the mass of one of the peaks in the spectrum is known, then a proportional scaling of the horizontal axis can be used to scale this peak, and thereby all other peaks, to the correct m/z-value. Alternatively, if none of the peak masses is known *a priori*, one can use our mixture model, which we describe in Section 2.4 to estimate the mass of the matrix adducts, μ_a. The adduct mass for the sinapinic acid (SPA) matrix is 206.06 Da, according to [1]. Therefore, multiplying the horizontal axis by a factor of $206.06/\mu_a$, should result in optimal scaling of the m/z-axis, too.

Figure 3 illustrates self-calibration in a spectrum with complex regions and many overlapping peaks. Figure 3A shows that self-calibration located the double- and triple-charge peaks at the correct relative position in the spectrum.

If two self-calibrated spectra (still) mis-align relative to each other, just a proportional scaling of the horizontal axis of one of the spectra will solve this. Our 64 spectra align well; we measured them in one batch, on the same day. Therefore, we self-calibrate the 64 spectra simultaneously by

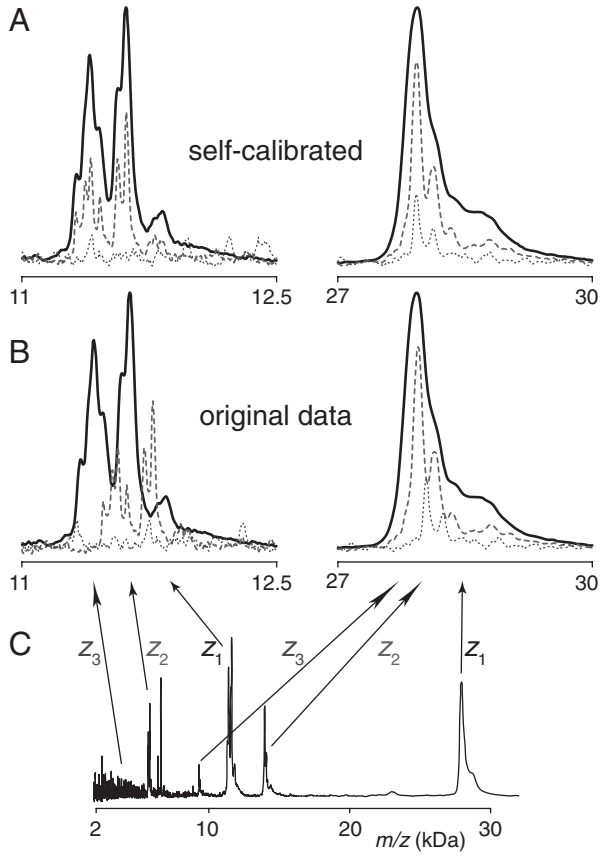

Figure 3. Detailed illustration of the self-calibration of spectrum (C). The solid black curves in (A) and (B) plot the z_1 peaks between 11–12.5 kDa (left) and between 27–30 kDa (right), in close-up. The z_2 (dashed curves) and z_3 (dotted curves) peaks are superimposed after multiplying their locations by 2 and 3, respectively. (B) shows that the relative locations of the z_1, z_2 and z_3 peaks in the original data do not exactly match. (A) shows that self-calibration matches the relative locations of the peaks.

optimizing their mean charge-correlations. Next, we scale all 64 *m/z*-axes simultaneously, as explained above. The exact trade-off of performing self-calibration on one spectrum *versus* on a bunch of similar ones together is open for further investigation.

2.4 Models interconnecting peak parameters

A molecule can form intermolecular complexes with other molecules. The matrix molecules, which are abundant in the SELDI analysis, frequently react with the molecules of interest by forming intermolecular complexes (particularly the SPA matrix). However, complexes between different sample molecules are less abundant and often generate peaks that do not exceed the noise level in the spectrum. We assume that a detected molecule can form non-covalent adducts (mass μ_a) with a matrix molecules, for $a \in 0,1,\ldots,a_{\max}$, where $a_{\max}$ is the maximum number of molecules reasonably involved in a single complex. In Fig. 1 we use SPA as matrix and we take $a_{\max} = 3$, because peaks containing four adducts do not exceed the noise level in the spectrum.

We assume that a detected molecule (or intermolecular complex) can carry z charges, for $z \in 1,2,\ldots,z_{\max}$, where $z_{\max}$ is the maximum number of charges that a molecule reasonably carries. For the analysis in Fig. 1, we take $z_{\max} = 3$, because peaks with four charges do not exceed the noise level in the spectrum.

In addition to different numbers of adducts and charges, isotopes also contribute to the multitude of peaks that can originate from a single molecular species. In SELDI data the resolution is generally too low to observe the individual isotopic peaks. Section 4 describes in detail how our models can be used and extended for the analysis of high-resolution data with isotopes.

We consider an experiment that consists of K spectra, numbered $k = 1,2,\ldots K$. Let $y_1, y_2, \ldots, y_I$ denote the *m/z*-values in the self-calibrated spectra and $n_{k,i}$ as the corresponding intensities, which correspond to the TOFs, $t_1, t_2, \ldots, t_I$, respectively, in spectrum k, where $y_i = y(t_i|\alpha,t_0,\beta)$. Suppose the sample contains M major molecular species, numbered $j = 1, 2, \ldots, M$, with molecular masses, μ_j. We assume that the *m/z*-values that are observed in a spectrum, k, derive from a mixture of a baseline distribution and $M \times (a_{\max}+1) \times z_{\max}$ normal distributions. In [1] we have shown that the peaks in the spectrum can be appropriately modeled with normal distributions. The normal distributions correspond to the observed peaks, and are defined by

$$f_{j,a,z}(y) = \frac{1}{\sigma_{j,a,z}\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_{j,a,z})^2}{2\sigma_{j,a,z}^2}\right) \quad (4)$$

where y is the observed *m/z*-value, the expected peak locations

$$\mu_{j,a,z} = \frac{\mu_j + a \cdot \mu_a}{z} \quad (5)$$

are the means of the distributions, and

$$\sigma_{j,a,z} = r \cdot \mu_{j,a,z}^2 \quad (6)$$

are the standard deviations of the distributions, for a parameter $r \in R_+$, which is related to the resolution of the peaks in the spectra.

We model the baseline ($f_{k,\mathrm{bl}}(y)$) with a lowess curve that uses locally weighted polynomial regression to fit the baseline in the data [17].

For spectrum k, the mixture distribution of the observed *m/z*-value y, is

$$f_k(y) = \sum_{j,a,z} p_{k,j,z} \cdot f_{j,a,z}(y) + p_{k,\mathrm{bl}} \cdot f_{k,\mathrm{bl}}(y) \quad (7)$$

where the parameters, $p_* \in R$ (*: any indexes) are the proportion parameters of the corresponding distributions, such that $0 \le p_*$, and such that the area under each mixture distribution equals 1, *i.e.* for each k, $\sum_{j,a,z} p_{k,*} = 1$. The

Supplementary Information, Section ''Parameter Estimation'' describes the parameter estimation and model visualization in detail.

3 Results

Figure 1 step by step (column by column) extends a simple mixture model to a more advanced mixture model. The peaks in this figure correspond to two detected molecular species in two spectra. The two species generate multiple (overlapping) peaks within one spectrum because of matrix adducts and multiple-charges. The upper two rows display the single-charge peaks in spectra 1 and 2. The third row displays the triple-charge peaks in spectrum 1. The first column shows the simple approach, one normal distribution *per* local mode in the data. The vertical lines (see left shaded rectangle) in columns 1 and 2 illustrate the discrepancies between the locations of the single-charge peaks and the expected locations of the triple-charge peaks. The second column incorporates the formation of matrix adducts in the model by adding an extra normal distribution for each matrix adduct. This improves the fit of the model to the data, and diminishes the discrepancies between the vertical lines. The third column links the parameters of peak components in our mixture models by making use of the known relationships between the locations of the peaks. Location estimations of corresponding peaks are linked across different spectra (rows 1 and 2), and within each spectrum (rows 1 and 3). Moreover, the parameters for the standard deviations are linked between all peaks in all spectra; *i.e.*, we only use one parameter (r) to model the standard deviations of all peaks. However, the goodness of fit is diminished in the third column. This is mainly because the spectra are not self-calibrated, or in other words, the triple-charge peak is not detected at 1/3 of the molecular mass of the single-charge peak. And, the double-charge peak is not detected at 1/2 of the mass of the single-charge peak (data not shown here). Therefore, we self-calibrate the spectra as illustrated in the fourth column.

In the 64 spectra, the peak shift due to self-calibration is on average 65 Da. These shifts are significant with respect to the instrument resolution: on average, a peak was shifted about 20 times its demi-width at half its maximum. We found that, as a consequence, self-calibration also improved the correlation between corresponding z_1 and z_2 peak sizes across the 64 spectra, from 0.37 in the original data to 0.53 after self-calibration. The z_1 and z_2 peak sizes of a given molecular species should have a high correlation across different spectra, because their ratio should be reproducible and independent of the molecular abundances in the different spectra.

The fourth column displays a parsimonious model (*i.e.* with a few parameters) that closely fits the data. We hereby reduce the total number of observed peaks to a much smaller number of underlying molecular species. Figure 4 shows that we can properly model the 29 952 peaks in our 64 spectra, using only 39 location parameters (μ_j). Moreover, we only use one parameter (r) to model the shapes (standard deviations) of all 29 952 peaks.

Figure 5 shows the fit of the parsimonious mixture model to another spectrum from the same data set. The right column (Cluster B) shows peaks in the same mass region as the peaks analyzed in Fig. 1. We have analyzed the single-charge peaks in Cluster A (shown in upper left plot) before in [13]. However, in that previous analysis we did not link the z_1 peaks to the corresponding z_2 peaks, as we do here. The z_2 peaks have higher resolution and help to better deconvolute the z_1 peaks. It is obvious that using the local maximum in the data will overestimate the peak size in the case of overlapping peaks. The upper left panel in Fig. 5 shows a clear example in which the local maximum is 74% higher than that of the underlying peak predicted by our model. We believe that the 96 peaks in this plot originate from only eight molecular species. Six of these species giving rise to the 72 peaks in Cluster A, and two giving rise to the 24 peaks in Cluster B. Other methods might not detect the peaks below the curly bracket in Cluster A, or, might explain these peaks as different molecules, *i.e.* independent from the other six molecules in Cluster A. As illustrated with the green peaks, our model can explain this complex region below the curly bracket by matrix adducts.

We even go a step further and make use of the adduct mass μ_a to come up with optimal m/z-values on the horizontal axis. The parameter μ_a is estimated after fitting our model to the data, and it should have a value of 206.06 Da, according to [1]. In our data set we estimate $\mu_a = 205.22$ Da. Using the known mass of the matrix adduct, we can now proportionally scale the m/z axis by a factor of $206.06/\mu_a$. This means we come up with m/z-values on the horizontal axis, purely on the basis of adduct formation and the combination of single- and double-charge peaks in the spectrum.

4 Discussion

In this article we developed novel methods and models for the optimal deconvolution analysis of MS data. We illustrated our models on complex and low-resolution MS (SELDI-TOF) data with commonly observed phenomena such as adduct formation and varying numbers of charges. Protein PTMs or degradations can also lead to interconnected peaks [18]. Our algorithm can take such interconnections into account, would the mass change of the modification be known *a priori* (information which is usually not available). We anticipate that our method and models have a general applicability to, and are very useful for, many commonly used MS separation, ionization and detection techniques.

Commonly used separation techniques that can be applied prior to MS analysis include LC and GC, capillary electrophoresis, IEF and 1-D gel electrophoresis and 2-DE. Our models can be used to link peaks across the different fractions that are separated by these techniques, in the same

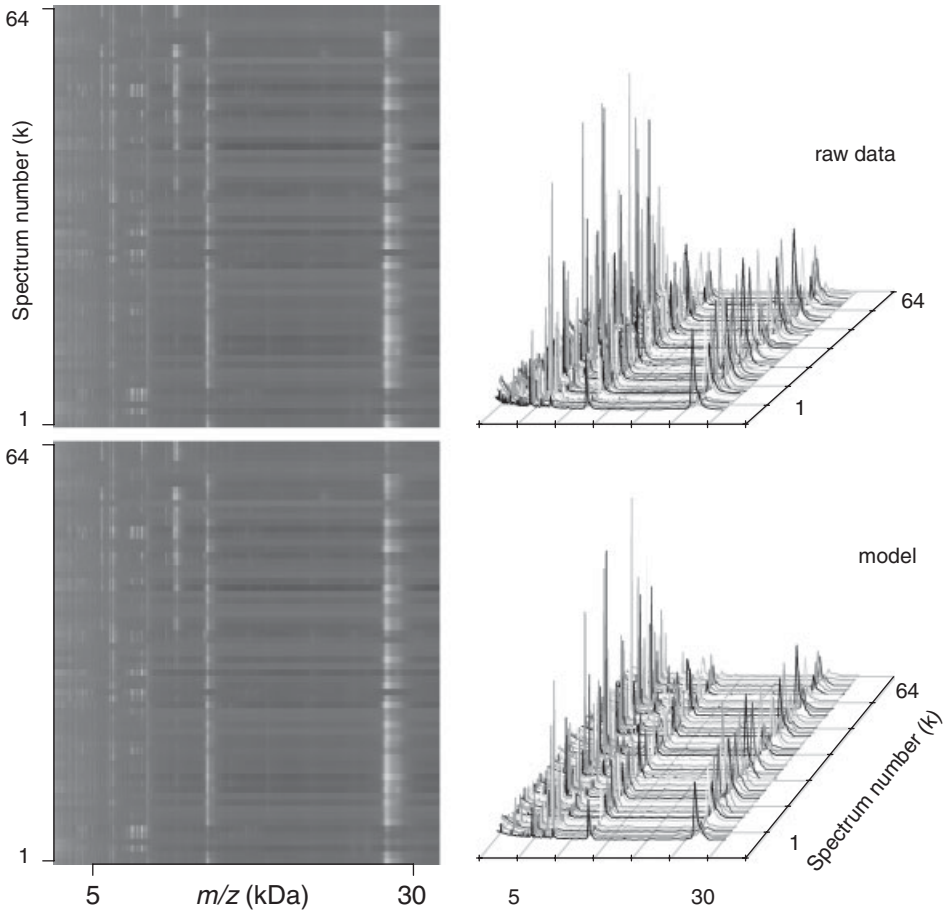

Figure 4. The upper two plots show the raw data of the 64 spectra in our experiment. The lower two plots show the 64 corresponding mixture models interconnecting peak parameters, which we fitted to the raw spectra. The left two plots show the data and the model as an image, where a white (black) shade corresponds to a high (low) intensity. The right two plots show data and model as graphs. Our models analyze the peaks of $M = 39$ molecular species. Because each molecule can get $z = 1$, 2 or 3 charges and/or $a = 0$, 1, 2 or 3 adducts, a spectrum contains $39 \times 3 \times 4 = 468$ peaks. In total, the 64 spectra contain $64 \times 468 = 29\,952$ peaks. Because our models link peaks, we only use 39 parameters to model the locations of all 29 952 peaks. Moreover, we only use one parameter (r) to model the shapes (standard deviations) of all 29 952 peaks. Using less parameters decreases the chances on overfitting and should thus lead to better estimates of molecular masses.

way as we link peaks between the different spectra in our experiment.

Besides MALDI, a commonly used ionization technique is ESI. With ESI, a molecule can get many more charges than with SELDI. We can take this into account by setting a higher value for z_{max}, *e.g.* 30.

Figure 2 illustrated that our current method for self-calibration optimizes the correlation between the intensities of z_1 peaks in region 1 and the intensities of the corresponding z_2 peaks in region 2. In addition to the considered peaks *per* region, other interrelated peaks may slightly contribute to the correlation, too. Next to z_1 peaks, region 1 may also contain z_2 peaks, and next to z_2 peaks, region 2 may also contain z_1 peaks. The z_2 peaks in region 1 contribute to the correlation because they correspond to z_4 peaks in region 2, and likewise z_1 peaks in region 2 correspond to dimers in region 1. A dimer consists of two molecules of the same species, which are linked together [1]. Z_4 Peaks and dimers generally have low relative abundances compared with z_1 and z_2 peaks, and may therefore only slightly contribute to the self-calibration. Absence of z_4 peaks and dimers is not expected to have a negative effect on the outcome of the self-calibration.

It is very unlikely that the z_1 peak of one given molecular species is matched to a peak of another species, and not to its z_2 peak, because of the following. Given calibration Eq. (1), matching one z_1 peak to a wrong location in the spectrum implies that all other z_1 peaks of other molecular species will also be matched to a wrong location in the spectrum and not to their corresponding z_2 peaks. This obviously is expected to lead to way lower correlation than a perfect match would do. Therefore, self-calibration is a robust approach to determine optimal values for the calibration parameters.

The Supplementary Information shows how our formula for the charge-correlation can be generalized. We anticipate that the generalized formula enables our methods to self-calibrate spectra in which molecules hold more charges (*e.g.* ESI), too.

Commonly used detection techniques are TOF, multipole, FT and orbitrap. These techniques can produce high-resolution spectra with peaks that show little or no overlap. Less overlap between peaks is favorable for the spectrum analysis because it simplifies the deconvolution analysis considerably. The authors of [19] analyzed Bovine Ubiquitin with ESI FT-MS and showed that the resolution of the resulting peaks was proportional to the charge on the molecule. This finding corresponds to our Eq. (6), which provides evidence that our parsimonious relationships between the standard deviations of our peaks also applies to ESI FT-MS data.

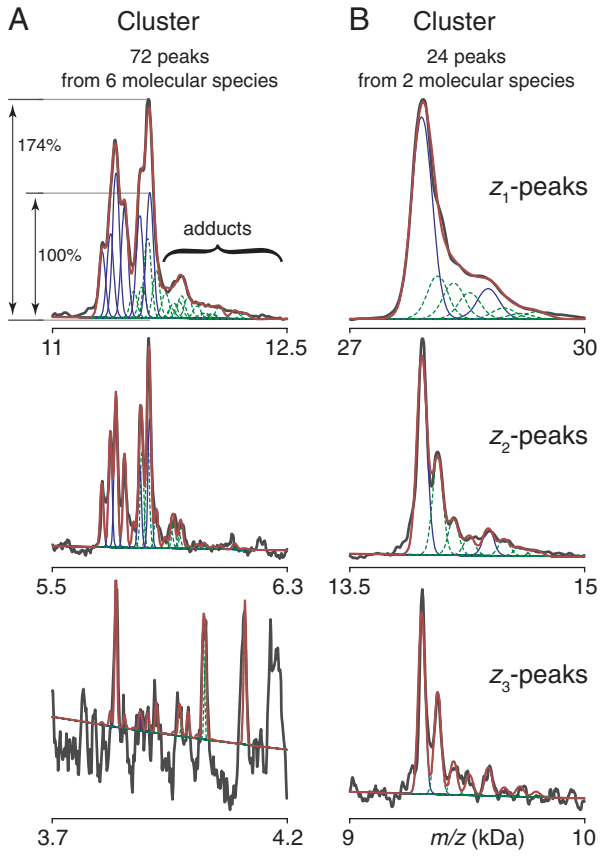

Figure 5. Deconvolution of two complex clusters with many overlapping peaks. Before deconvolution, the spectrum was self-calibrated, as shown in Fig. 3. Rows 1, 2 and 3 show the single-, double- and triple-charge peaks, respectively. The data are plotted in black. The fitted mixture distribution is plotted in red. The thin blue and dashed green curves indicate the individual mixture components; blue means 0 adducts, green means >0 adducts. We expect that the peaks below the curly bracket originate from matrix adducts. The local maximum indicated in the upper left plot is 74% higher than the underlying peak as predicted by our models.

High-resolution metabolomics and proteomics spectra can detect isotopes and common chemical transformations [18]. We imply that it is straightforward to incorporate these phenomena parsimoniously in our models.

We anticipate that, in protein spectra, we may moreover define parsimonious interrelationships between the proportions of isotopic peaks. Proteins consist of amino acids. An interesting property of amino acids is that their composition is mainly limited to the chemical elements carbon (C), hydrogen (H), nitrogen (N), oxygen (O) and sulfur (S) atoms. Based on the average amino acid composition, Senko *et al.* derived a model amino acid "averagine" [20]. The molecular formula of averagine:

$$C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$$

Senko *et al.* can accurately predict the isotopic distribution of proteins of each given mass, based on this average protein composition. We can use the "average protein composition" to incorporate the predicted isotopic distribution in our model. This will result in parsimonious models with interconnected proportions of interrelated isotopes, which, we imply, are very suitable for the analysis of high-resolution spectra with peaks on the isotopic mass level.

Optimization of the correlation between peaks with a known mass difference, as function of the calibration parameters, should improve the self-calibration of a given spectrum. In a similar way, we imply that the regular distances between isotopes in high-resolution spectra can be used to further improve the self-calibration.

We imply that our novel models help to improve biomarker discovery for the following reasons. We can detect peaks in complex regions of the spectrum since we make use of information from related regions with lower complexity and higher resolution, by linking peaks. This is important because each peak is a potential biomarker. Moreover, we produce appropriate estimates of the peak positions and the molecule masses. These estimates can help in subsequent (biomarker) molecule identification steps. We also improve the estimates of the molecule abundances, which increases the chance on finding "real" biomarkers in the discovery phase. An additional improvement for biomarker discovery is that by linking peaks we reduce the total number of observed peaks in a spectrum to a much smaller number of underlying molecular species. This reduces the statistical test multiplicity in the biomarker discovery phase and therefore increases the power, and ultimately the chance on finding real biomarkers even further.

5 Concluding remarks

In this paper, we presented a novel method, called "*self-calibration*", to locate peaks at the correct locations in the spectrum. Self-calibration can be applied to any spectrum, even if the sample content is unknown and when the original TOFs and calibration parameters are not available. Moreover, we implied that our novel statistical models linking peaks have a wide applicability to commonly used MS techniques, improve biomarker discovery and have better power to get more out of your MS data.

NGI / NBIC / BioAssist supported this work.

The authors have declared no conflict of interest.

6 References

[1] Dijkstra, M., Vonk, R. J., Jansen, R. C., SELDI-TOF mass spectra: a view on sources of variation. *Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, *847*, 12–23.

[2] Tan, C. S., Ploner, A., Quandt, A., Lehtiö, J. *et al.*, Annotated regions of significance of SELDI-TOF-MS spectra for detecting protein biomarkers. *Proteomics* 2006, *6*, 6124–6133.

[3] Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A. *et al.*, Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* 2005, *5*, 4107–4117.

[4] Steffen, B., Müller, K. P., Komenda, M., Koppmann, R., Schaub, A., A new mathematical procedure to evaluate peaks in complex chromatograms. *J. Chromatogr. A* 2005, *1071*, 239–246.

[5] Vivó-Truyols, G., Torres-Lapasió, J. R., van Nederkassel, A. M., Heyden, Y. V., Massart, D. L., Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: peak detection. *J. Chromatogr. A* 2005, *1096*, 133–145.

[6] Vivó-Truyols, G., Torres-Lapasió, J. R., van Nederkassel, A. M., Heyden, Y. V., Massart, D. L., Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part II: peak model and deconvolution algorithms. *J. Chromatogr. A* 2005, *1096*, 146–155.

[7] Carlson, S. M., Najmi, A., Whitin, J. C., Cohen, H. J., Improving feature detection and analysis of surface-enhanced laser desorption/ionization-time of flight mass spectra. *Proteomics* 2005, *5*, 2778–2788.

[8] Hu, J., Coombes, K. R., Morris, J. S., Baggerly, K. A., The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief. Funct. Genomic. Proteomic.* 2005, *3*, 322–331.

[9] Altamura, S., Kiss, J., Blattmann, C., Gilles, W., Muckenthaler, M. U., Seldi-tof ms detection of urinary hepcidin. *Biochimie* 2009. doi:10.1016/j.biochi.2009.04.010.

[10] Melle, C., Ernst, G., Winkler, R., Schimmel, B. *et al.*, Proteomic analysis of human papillomavirus-related oral squamous cell carcinoma: identification of thioredoxin and epidermal-fatty acid binding protein as upregulated protein markers in microdissected tumor tissue. *Proteomics* 2009, *9*, 2193–2201.

[11] Calvo, F. Q., Fillet, M., de Seny, D., Meuwis, M.-A. *et al.*, Biomarker discovery in asthma-related inflammation and remodeling. *Proteomics* 2009, *9*, 2163–2170.

[12] Emanuele, V. A., Gurbaxani, B. M., Benchmarking currently available seldi-tof ms preprocessing techniques. *Proteomics* 2009, *9*, 1754–1762.

[13] Dijkstra, M., Roelofsen, H., Vonk, R. J., Jansen, R. C., Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics* 2006, *6*, 5106–5116.

[14] Roelofsen, H., Alvarez-Llamas, G., Dijkstra, M., Breitling, R. *et al.*, Analyses of intricate kinetics of the serum proteome during and after colon surgery by protein expression time series. *Proteomics* 2007, *7*, 3219–3228.

[15] Jeffries, N., Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics* 2005, *21*, 3066–3073.

[16] Ciphergen Biosystems Inc., Proteinchip Software 3.1 *Operation Manual* 2002.

[17] Cleveland, W. S., Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 1979, *74*, 829–836.

[18] Breitling, R., Pitt, A. R., Barrett, M. P., Precision mapping of the metabolome. *Trends Biotechnol.* 2006, *24*, 543–548.

[19] Marshall, A. G., Hendrickson, C. L., Charge reduction lowers mass resolving power for isotopically resolved electrospray ionization fourier transform ion cyclotron resonance mass spectra. *Rapid commun. Mass Spectrom.* 2001, *15*, 232–235.

[20] Senko, M., Beu, S., McLafferty, F., Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 1995, *6*, 229–233.

[21] Yasui, Y., Pepe, M., Thompson, M. L., Adam, B.-L. *et al.*, A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2003, *4*, 449–463.

[22] de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O., *Computational Geometry*, Springer, Germany 2000.

[23] Dempster, A., Laird, D., Rubin, J., Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B* 1977, *39*, 1–38.

[24] Tierney, L., Rossini, A. J., Li, N., Snow: a parallel computing framework for the r system. *Int. J. Parallel Program.* 2009, *37*, 78–90.

